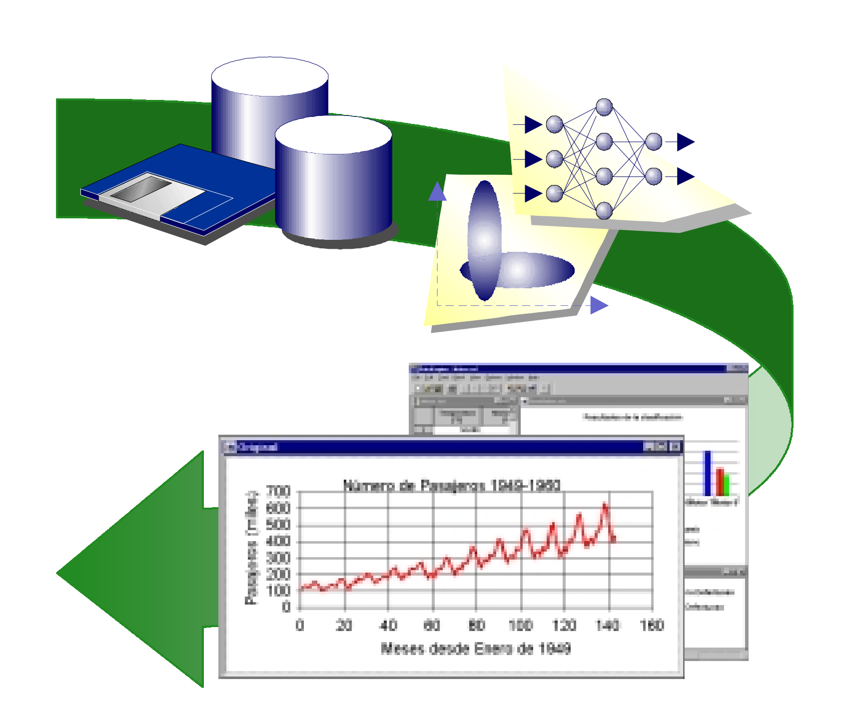


MINERÍA DE DATOS

Conceptos y Objetivos



Documento Básico

1012-V-DB-002-020 - Septiembre de 2000
DAEDALUS – Data, Decisions and Language, S.A.
http://www.daedalus.es

Por favor envíen sus sugerencias y comentarios a:

DAEDALUS – Data, Decisions and Language, S.A.
C/ Pirineos, 27, 1º
E-28040 Madrid

Tel: +34 91 311 33 86
Fax: +34 91 450 40 58
email:info@daedalus.es
http://www.daedalus.es



ÍNDICE

MINERÍA DE DATOS

Documento básico

1 RESUMEN

Con la denominada sociedad de la información se está produciendo un fenómeno curioso. Día a día se multiplica la cantidad de datos almacenados. Sin embargo, contrariamente a lo que pudiera esperarse, esta explosión de datos no supone un aumento de nuestro conocimiento, puesto que resulta imposible procesarlos con los métodos clásicos. La mayoría de las multinacionales generan más información en una semana que la que cualquier persona podría leer en toda su vida, e incluso las pequeñas empresas generan un volumen de datos que no son capaces de manejar. De modo que actualmente nos enfrentamos a la paradoja de que, cuantos más datos están disponibles, menos información tenemos.

Para superar este problema, en los últimos años han surgido una serie de técnicas que facilitan el procesamiento avanzado de los datos y permiten realizar un análisis en profundidad de los mismos de forma automática. La idea clave es que los datos contienen más información oculta de la que se ve a simple vista.

Este documento ofrece una perspectiva general del proceso completo de extracción del conocimiento oculto en los datos, denominado *KDD* (*Knowlegde Discovery in Databases*) y, más en concreto, de las técnicas utilizadas en la fase de descubrimiento de información propiamente dicha, denominada minería de datos.

En la sección 2 se definen una serie de conceptos básicos que ayudarán a situar la minería de datos dentro de diferentes entornos de trabajo. Además se definen conceptos asociados a la minería de datos como *Data Warehouse* y OLAP.

Si el lector ya tiene una cierta idea de lo que se puede conseguir con la minería de datos, y lo que desea es saber qué tipo de aplicaciones o problemas pueden ser abordados con esta tecnología, puede ir directamente a las sección 3. En ella se establecen los objetivos que podemos alcanzar con las diferentes técnicas existentes.

Este documento puede complementarse con otros documentos básicos accesibles en www.daedalus.es:

- **Minería de datos – Tecnología**, en el que se profundiza en las técnicas propias de la minería de datos.
- **Desarrollo de proyectos de minería de datos**, donde se describe la metodología CRISP-DM, un estándar industrial con más de 160 empresas y organizaciones en su grupo de interés.
- **Web Mining – Minería de uso de la Web**. Si su trabajo se desarrolla en el mundo de Internet, seguro que le interesará cómo aplicar la minería de datos a este ámbito.



2 CLAVES OCULTAS EN SUS DATOS

2.1 Los datos, origen de la información

Hoy en día, y está claro que se trata de una tendencia válida para los próximos años, el almacenamiento de la información es algo sencillo y barato. Nuestros sistemas informáticos cada vez tienen una capacidad mayor, y lo que ahora es normal encontrar "de serie" en un ordenador personal, quedará anticuado dentro de unos meses.

Este incremento de los sistemas de almacenamiento tiene un efecto que es realmente interesante: es poco costoso guardar datos del funcionamiento de nuestros procesos, o de nuestros sistemas de venta, o de nuestros clientes, etc., por lo que nuestras bases de datos (en el sentido más amplio del término) crecen hasta límites insospechados.

Cuando decidimos iniciar ese proceso de almacenamiento de datos, lo solemos hacer con la intención de analizarlos posteriormente. Sin embargo, cuando llega el momento, el análisis que se realiza suele ser bastante superficial y guiado por los resultados que esperamos encontrar al analizarlos. Lo normal es utilizar algún paquete estadístico (una hoja de cálculo en el caso más simple) para localizar correlaciones entre variables, establecer medias y varianzas e intentar modelar de esta forma nuestra información.

Sin embargo, en esa montaña de datos existe información que no puede ser encontrada con los procedimientos habituales de trabajo. La minería de datos nos ayuda a dar un paso más en ese análisis sacando a la luz relaciones ocultas entre los datos: información desconocida que pueda ayudarnos a gestionar mejor nuestro negocio o proceso.

2.2 Estructuración de los datos

Para poder analizar nuestros datos con fiabilidad es necesario que exista una cierta estructuración y coherencia entre los mismos. Si el responsable de almacenamiento de la información ha sido siempre la misma persona, es posible que una parte de este problema esté resuelto. Sin embargo, en general no se da esa situación, sino que, más bien al contrario, son muchas las personas que en distintos departamentos y a lo largo del tiempo han ido creando ficheros con diferentes tipos de datos.

Surge aquí la necesidad de conjugar los distintos ficheros y bases de datos de manera que podamos utilizarlos para extraer conclusiones. Aunque más adelante trataremos el problema del preprocesamiento de los datos, en este punto podemos echar un vistazo a los problemas que podemos encontrarnos:

- Diferentes tipos de datos representando el mismo concepto: un ejemplo que ha provocado uno de los mayores problemas informáticos es la representación de la fecha, donde el año se puede guardar con 2 o con 4 dígitos.
- Diferentes claves para representar el mismo elemento: un mismo cliente puede ser representado por un código de cliente propio o por su NIF.
- Diferentes niveles de precisión al representar un dato: los números reales no siempre se almacenan de la misma forma, y es posible que esto nos genere algún problema.
- ...

Como podemos ver, la cuestión no es sencilla, y se agrava cuando los diferentes ficheros se encuentran en sistemas informáticos y soportes diferentes.

Es cierto que cada una de estas fuentes de datos puede ser manejada por separado. Seguro que hay quien opina que los datos están en diferentes ficheros porque representan informaciones y procesos distintos, y que no tiene sentido estructurar la información más allá de lo que ya está. Y es posible que si así lo hacemos encontremos información útil. Pero no es menos cierto que nos estamos hurtando a nosotros mismos la posibilidad de descubrir un conocimiento que va más allá de cada una de las parcelas de nuestro negocio: un conocimiento que representa la interacción entre diferentes procesos, que es, precisamente, donde se encuentra la información más valiosa.

2.3 Data Warehousing

El mecanismo más habitual para estructurar la información de un negocio es haciendo uso de un *Data Warehouse*[1]. Las definiciones más habituales de este término son:

- **Almacén de datos**. Plataforma que concentra la información de interés de toda la empresa.
- Sistema que permite el almacenamiento en un único entorno de la **información histórica e integrada** proveniente de los distintos sistemas de la empresa y que **refleja los indicadores clave** asociados a los negocios de la misma.
- Sistema de información **orientado a la toma de decisiones** empresariales que, almacenando de **manera integrada** la información relevante del negocio, permite la realización de **consultas complejas con tiempos de respuesta cortos**.
- Sistema orientado a dar **información en términos de negocio** en vez de datos en términos de explotación.

Como se puede apreciar, las palabras más empleadas son: información de interés, negocio, integración,... De su conjunto podemos expresar que el *Data Warehouse* es un almacén estructurado de la información clave de nuestro negocio, que integra datos provenientes de todos los departamentos, sistemas, etc. y que nos permite analizar el funcionamiento de nuestra compañía y tomar decisiones sobre su gestión.

No se trata de una simple agregación de las diferentes bases de datos. Es importante destacar que hay algunas diferencias de concepto respecto a éstas y a su forma de uso.

Una base de datos operativa almacena la información de un sector del negocio, se actualiza a medida que llegan datos que deban ser almacenados y se opera mediante los cuatro mecanismos clásicos "Añadir-Eliminar-Modificar-Imprimir":

- Clásicamente se orienta hacia la elaboración de informes periódicos.
- Suele manejar pequeños volúmenes de datos.
- Entorno dimensionado para muchas transacciones (gran cantidad de actualizaciones).

En cuanto al *Data Warehouse*, su actualización se realiza a intervalos regulares (típicamente una al día) dentro de un proceso controlado, y tras realizar un preprocesado de los datos que se van a almacenar. Su orientación es hacia la consulta del estado del negocio.

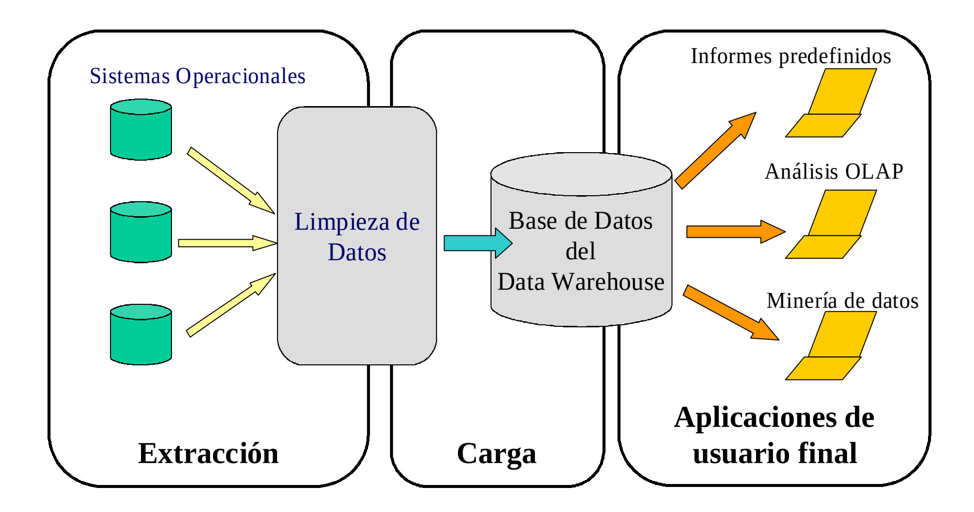
- Se ofrece información bajo demanda (análisis libre mediante el uso de herramientas de generación de informes que atacan el *Data Warehouse*).
- Refleja el modelo de negocio, frente al modelo de proceso.

---

[1] El término *Data Warehouse* es de difícil traducción como una sola palabra. No existe un término único aceptado comúnmente, por lo que hemos preferido mantenerlo en su idioma original.



- Almacena grandes volúmenes de datos (información histórica e integración de datos de múltiples aplicaciones).
- Dimensionado para consultas largas y elaboradas.
- Actualizaciones controladas y no eliminación de datos (el *Data Warehouse* contiene toda la historia de la compañía).



La estructura de esta gran base de datos es multidimensional, con diferentes puntos de vista que reflejan los distintos aspectos del negocio. Así los responsables de producto pueden analizar su evolución a lo largo del tiempo en diferentes sectores y localización geográfica. Sobre los mismos datos, los responsables de grandes cuentas pueden obtener información sobre los tipos de productos que se han vendido, por regiones, a lo largo del tiempo. Un director regional podrá estudiar cómo evoluciona su mercado particular, etc.

El ejemplo clásico para representar un *Data Warehouse* es el de un cubo de datos, del que se pueden extraer diferentes "rodajas" o puntos de vista, se puede analizar una parte concreta, o estudiar el conjunto global. Más adelante, cuando describamos las herramientas OLAP, volveremos sobre esta idea.

Cuando mantenemos una estructura de *Data Warehouse*, pero adaptada sólo a un sector de la empresa, o para un fin concreto, se utiliza un *Data Mart*. Los *Data Marts* pueden extraerse del *Data Warehouse* de la empresa, aunque también es posible que el *Data Warehouse* se construya a partir de los *Data Marts* que se hayan ido diseñando e implantando en los diferentes departamentos. Este segundo enfoque es el que se utiliza cuando se comienza por aplicar estas técnicas en algunas de las áreas del negocio y no en su globalidad.

2.4 Información oculta en los datos

A estas alturas ya va pareciendo claro que si almacenamos la información más relevante de nuestro negocio en un sistema que acumula y acumula datos sin parar, un análisis razonable nos puede permitir descubrir tendencias, localizar grupos de datos con comportamiento homogéneo, establecer relaciones, etc.

Esa información está oculta en los datos y será necesario utilizar todas las técnicas a nuestro alcance para obtenerla. El objetivo que nos planteamos es localizar relaciones entre atributos de nuestro *Data Warehouse*. Estas relaciones podrían ser del tipo:

- *Para una gran superficie*: Más del 60% de las personas que adquieren queso fresco compran también algún tipo de mermelada.

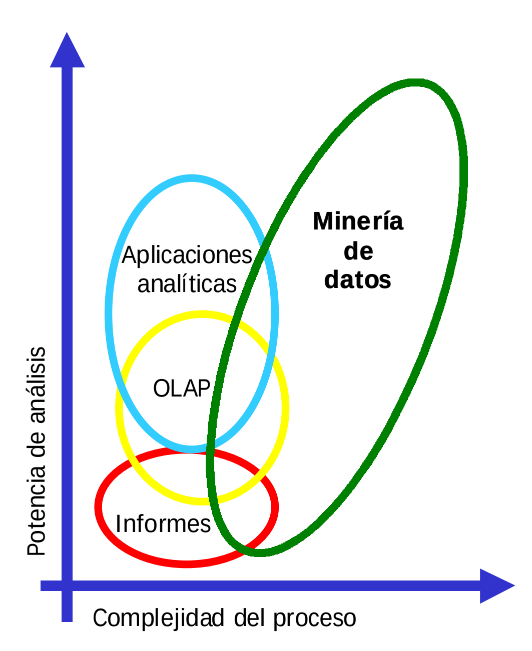
- *Para un departamento de fidelización de una compañía aérea*: muchos usuarios que hacen vuelos de menos de 3 días a Berlín alquilan un coche en el aeropuerto.
- *Para un operador de telefonía*: durante el mes siguiente al lanzamiento de una campaña de descuento en llamadas internacionales por parte de una compañía de la competencia, nuestros pequeños clientes redujeron su consumo en este sector, mientras que los grandes clientes lo mantuvieron.

Esta información puede ser extraída haciendo uso de diversas técnicas y ninguna de ellas debe ser despreciada, sino agregada al resto para obtener mejores resultados. Sin embargo, en este documento básico nos centraremos en la minería de datos y en las ventajas que puede aportar frente a otras técnicas.

2.5 Qué es y qué no es la minería de datos

La minería de datos puede definirse como la **extracción no trivial de información implícita, previamente desconocida y potencialmente útil, a partir de los datos**[2]. Para conseguirlo hace uso de diferentes tecnologías que resuelven problemas típicos de agrupamiento automático, clasificación, asociación de atributos y detección de patrones secuenciales. La minería de datos es, en principio, una fase dentro de un proceso global denominado **descubrimiento de conocimiento en bases de datos** (*Knowledge Discovery in Databases* o *KDD*), aunque finalmente haya adquirido el significado de todo el proceso[3] en lugar de la fase de extracción de conocimiento.

Es habitual que los expertos en estadística confundan la minería de datos con un análisis estadístico de éstos (afirmaciones de este tipo pueden encontrarse en documentación de empresas dedicadas al procesamiento estadístico que *venden* sus productos como herramientas de minería de datos). La diferencia fundamental entre ambas técnicas es muy clara: para conseguir una afirmación como la que ha sido utilizada en el ejemplo anterior (*Más del 60% de las personas que adquieren queso fresco compran también algún tipo de mermelada)* utilizando un paquete estadístico, es necesario conocer a priori que existe una relación entre el queso fresco y la mermelada, y lo que realizamos con nuestro entorno estadístico es una cuantificación de dicha relación.



En el caso de la minería de datos el proceso es muy distinto: la consulta que se realiza a la base de datos (al *Data Warehouse*) busca relaciones entre parejas de productos que son adquiridos por una misma persona en una misma compra. De esa información, el sistema deduce, junto a otras muchas, la afirmación anterior. Como podemos ver, en este proceso se realiza un acto de descubrimiento de conocimiento real, puesto que no es necesario ni siquiera sospechar la existencia de una relación entre estos dos productos para encontrarla.

---

[2] W. Frawley, G. Piatesky-Shapiro, C. Matheus, “Knowledge Discovery in Databases: An Overview”, AI Magazine, Otoño 1992 (pág. 213-228).

[3] En la mayoría de la bibliografía se hace referencia a minería de datos tomando el sentido de “descubrimiento de conocimiento en bases de datos”.



2.6 Definición, caracterización y estructura del problema

La evolución de la tecnología ha facilitado y automatizado en gran medida las tareas de análisis de información. Cada paso en esta evolución se apoya en los anteriores y cada uno de ellos ha supuesto un avance significativo para el usuario, que ha visto cómo cada progreso le abría nuevas posibilidades de análisis y aumentaba el nivel de abstracción de las consultas.

Para decidir cuál es la técnica más adecuada para una determinada situación, es necesario distinguir el tipo de información que se desea extraer de los datos. Según su nivel de abstracción, el conocimiento contenido en los datos puede clasificarse en distintas categorías y requerirá una técnica más o menos avanzada para su recuperación:

- ***Conocimiento evidente***

  Información fácilmente recuperable con una simple consulta (SQL). Un ejemplo de este tipo de conocimiento es una pregunta como "¿Cuáles fueron las ventas en España el pasado marzo?" o "¿Cuál es la edad media de mis clientes?".

- ***Conocimiento multi-dimensional***

  El siguiente nivel de abstracción consiste en considerar los datos con una cierta estructura. Por ejemplo, en vez de considerar cada transacción individualmente, las ventas de una compañía pueden organizarse en función del tiempo y de la zona geográfica, y analizarse con diferentes niveles de detalle (país, región, localidad...).
  Técnicamente, se trata de reinterpretar una tabla con *n* atributos independientes como un espacio *n*-dimensional, lo que permite detectar algunas regularidades difíciles de observar con la representación monodimensional clásica. Este tipo de información es la que analizan las herramientas OLAP, que resuelven de forma automática cuestiones como "¿Cuáles fueron las ventas en España el pasado marzo? Aumentar el nivel de detalle: mostrar las de Madrid."

- ***Conocimiento oculto***

  Información no evidente, desconocida a priori y potencialmente útil, que puede recuperarse mediante técnicas de minería de datos, como reconocimiento de regularidades. Esta información es de gran valor, puesto que no se conocía y se trata de un descubrimiento real de nuevo conocimiento, del que antes no se tenía idea, y que abre una nueva visión del problema. Un ejemplo de este tipo sería "¿Qué tipos de clientes tenemos? ¿Cuál es el perfil típico de cada clase de usuario?".

Como se ve, las técnicas disponibles para extraer la información contenida en los datos son muy variadas y cada una de ellas es complementaria del resto, no exclusivas entre sí. Cada técnica resuelve problemas de determinadas características y, para extraer todo el conocimiento oculto, en general será necesario utilizar una combinación de varias.

La mayor parte de la información de interés contenida en una base de datos, aproximadamente el 80%, corresponde a conocimiento superficial, fácilmente recuperable mediante consultas sencillas con SQL. El 20% restante corresponde a conocimiento oculto que requiere técnicas más avanzadas de análisis para su recuperación. Estas cifras pueden dar la falsa impresión de que la cantidad de información recuperable mediante técnicas de minería de datos es despreciable. Sin

embargo, se trata precisamente de información que puede resultar de vital importancia para la empresa y que no se puede desdeñar.

Básicamente, y como ya hemos comentado, la clave que diferencia la minería de datos respecto de las técnicas clásicas es que el análisis que realiza es *exploratorio*, no *corroborativo*. Se trata de descubrir conocimiento nuevo, no de confirmar o desmentir hipótesis. Con cualquiera de las otras técnicas es necesario tener una idea concreta de lo que se está buscando y, por tanto, la información que se obtiene con ellas está condicionada a la idea preconcebida con que se aborde el problema. Con la minería de datos es el sistema y no el usuario el que encuentra las hipótesis, además de comprobar su validez.

La minería de datos, esencialmente, permite obtener a partir de los datos un **modelo** del problema que se analiza, bien sean las ventas de un artículo para mejorar la campaña de marketing, las características técnicas de un producto en control de calidad o un proceso industrial cuyo control se desea optimizar, por citar algunos ejemplos. El modelo obtenido permitirá simular el comportamiento del sistema real y obtener conclusiones aplicables en el día a día.

2.7 ¿Estamos dispuestos a usar los resultados?

La minería de datos descubre relaciones en los datos, pero eso es sólo el principio. Son las personas, no las técnicas de minería de datos, las que toman decisiones. El factor más importante en minería de datos es el conocimiento y la experiencia de dichas personas. Armadas con información mejor, pueden aplicar su creatividad y su propio criterio para tomar decisiones más acertadas y obtener mejores resultados.

Por muy buenos que sean los resultados obtenidos en un proyecto de minería de datos, son totalmente inútiles si no se aplican en la práctica. Así, es inútil que consigamos un clasificador que diferencie perfectamente diversos tipos de clientes si no se tiene en cuenta dicha información en una campaña de marketing. O descubrir la influencia de una determinada variable en el rendimiento de un proceso si después no se controla consecuentemente su valor. Las conclusiones de la minería de datos no son valiosas por sí mismas, sino en la medida en que se apliquen para obtener resultados.

Es importante recordar que los responsables de dicha puesta en práctica no serán generalmente expertos en minería de datos. Un factor clave en el éxito de estos proyectos es presentar los resultados de una forma clara e inteligible, haciendo hincapié en la información realmente útil, teniendo siempre en cuenta sus destinatarios. Es asimismo fundamental justificar adecuadamente dichas conclusiones, puesto que otro problema muy generalizado es la desconfianza que frecuentemente suscitan los sistemas automáticos. A menudo, es necesario un cambio de mentalidad para convencer a las personas involucradas del interés, utilidad y fiabilidad de la información obtenida gracias a la minería de datos.

Estas dificultades pueden ser superadas en gran medida si los responsables de la aplicación del proyecto han participado activamente en su desarrollo. Será mucho más sencillo convencer a una persona de la validez de las conclusiones obtenidas si ella misma ha aportado su conocimiento del proceso en estudio, o de su utilidad si fue el promotor del análisis. La colaboración de todos los usuarios implicados es fundamental para el éxito de un proyecto de minería de datos.



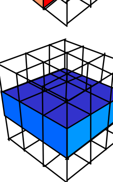
3 PARA QUÉ SIRVE LA MINERÍA DE DATOS

3.1 *Minería de datos frente a OLAP y DSS*

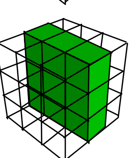
Los sistemas de ayuda a la decisión (*DSS*) son herramientas sobre las que se apoyan los responsables de una empresa, directivos y gestores, en la toma de decisiones. Para ello, utilizan:

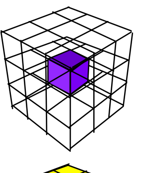
- un *Data Warehouse*, en el que se almacena la información de interés para la empresa, y
- herramientas de análisis multidimensional (OLAP).

OLAP (On-Line Analytical Processing) se define como análisis rápido de información multidimensional compartida[4]. El término OLAP aparece en contraposición al concepto tradicional OLTP (On-Line Transactional Processing), que designa el procesamiento *operacional* de los datos, orientado a conseguir la máxima eficacia y rapidez en las transacciones (actualizaciones) individuales de los datos, y no a su análisis de forma agregada.

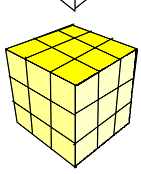


Las herramientas OLAP permiten navegar a través de los datos almacenados en el *Data Warehouse* y analizarlos dinámicamente desde una perspectiva multidimensional, es decir, considerando unas variables en relación con otras y no de forma independiente entre sí y permitiendo enfocar el análisis desde distintos puntos de vista. Esta visión multidimensional de los datos puede visualizarse como un "cubo de Rubik", que puede girarse para examinarlo desde distintos puntos de vista, y del que se pueden seleccionar distintas "rodajas" o "cubos" dependiendo de los aspectos de interés para el análisis.





Los DSS permiten al responsable de la toma de decisiones consultar y utilizar de manera rápida y económica las enormes cantidades de datos operacionales y de mercado que se generan en una empresa. Gracias al análisis OLAP, pueden verificarse hipótesis y resolverse consultas complejas. Además, en el curso del análisis, la interpretación de los datos puede dar lugar a nuevas ideas y enfoques del problema, sugiriendo nuevas posibilidades de análisis.



Sin embargo, el análisis OLAP depende de un usuario que plantee una consulta o hipótesis. Es el usuario el que lo dirige y, por tanto, el análisis queda limitado por las ideas preconcebidas que aquél pueda tener.

La minería de datos constituye un paso más en el análisis de los datos de la empresa para apoyar la toma de decisiones. No se trata de una técnica que sustituya los DSS ni el análisis OLAP, sino que los complementa, permitiendo realizar un análisis más avanzado de los datos y extraer más información de ellos.

Como ya se ha comentado anteriormente, utilizando minería de datos es el propio sistema el que descubre nuevas hipótesis y relaciones. De este modo, el conocimiento obtenido con estas técnicas no queda limitado por la visión que el usuario tiene del problema.

Las diferencias entre minería de datos y OLAP radican esencialmente en que el enfoque desde el que se aborda el análisis con cada una de ellas es completamente distinto. Fundamentalmente:

[4] Richard Creeth, Nigel Pendse.

- El análisis que realizan las herramientas OLAP es dirigido por el usuario, deductivo, parte de una hipótesis o de una pregunta del usuario y se analizan los datos para resolver esa consulta concreta. Por el contrario, la minería de datos permite razonar de forma inductiva a partir de los datos para llegar a una hipótesis general que modele el problema.

- Además, las aplicaciones OLAP trabajan generalmente con datos agregados, para obtener una visión global del negocio. Por el contrario, la minería de datos trabaja con datos individuales, concretos, descubriendo las regularidades y patrones que presentan entre sí y generalizando a partir de ellos.

	OLAP	**MD**
Razonamiento	deductivo	inductivo
Trabaja con datos	agregados	concretos/individuales

Un ejemplo clarificará la diferencia entre ambas técnicas:

Una pregunta típica de un sistema OLAP/DSS sería: “El año pasado, ¿se compraron más furgonetas en Cataluña o en Madrid?”. La respuesta del sistema sería del tipo “En Cataluña se compraron 12.000 furgonetas, mientras que, durante el mismo intervalo, en Madrid se compraron 10.000”. Obviamente es una información interesante y útil, pero restringida por las hipótesis realizadas a priori.

En cambio, un problema típico para resolver utilizando minería de datos sería, por ejemplo: “Hallar un modelo que determine las características más relevantes de las personas que compran furgonetas”. A partir de los datos del pasado, el sistema de minería de datos proporcionaría una respuesta del tipo: “Depende de la época del año y la situación geográfica. En invierno, los habitantes de Madrid que pertenecen a un cierto grupo de edad y nivel de ingresos probablemente comprarán más furgonetas que gente de las mismas características en Cataluña”.

Como puede verse, se trata de problemas distintos, de modo que según los objetivos perseguidos deberá utilizarse una técnica u otra. Además, puesto que sus conclusiones son complementarias, en general será conveniente combinar ambas para obtener los mejores resultados.





3.2 ¿Qué se puede esperar?

El objetivo final de cualquier proyecto de minería de datos puede resumirse en uno de estos dos:

- ✓ ahorrar dinero mejorando la eficacia de sus actividades, o bien,
- ✓ ganar dinero descubriendo nuevas fuentes de beneficios.

¿Cómo se llega a estos objetivos? A partir de un conjunto de datos y un conjunto de técnicas se puede llegar a unas determinadas conclusiones. Pero, ¿cómo se traducen los resultados de un proyecto de minería de datos en beneficios tangibles para la empresa? Básicamente, esos resultados suponen una mejora de la información disponible y será al aplicar dicha información cuando se obtengan los beneficios.

Los campos en los que pueden utilizarse estas técnicas son extremadamente variados: prácticamente en cualquier situación en la que se disponga de un conjunto de datos. A continuación se comentan algunas de las áreas más comunes en las que se ha aplicado frecuentemente la minería de datos, pero se trata simplemente de algunos ejemplos. En casi cualquier caso que usted pueda imaginar es probable que la minería de datos pueda aportar importantes beneficios.

¿Parece una exageración? Tal vez no tanto. A modo de curiosidad: 28 de los 29 equipos que participan en la liga de baloncesto profesional americana (NBA) utilizan técnicas de minería de datos para detectar patrones de comportamiento y relaciones entre variables del juego (por ejemplo, detectar que el jugador X realiza el 90% de sus tiros de campo cuando el jugador Y juega de base), de forma que estas técnicas ofrecen nuevas perspectivas para modificar las tácticas de juego a fin de mejorar el rendimiento del equipo. Un análisis tradicional podría indicar que un jugador consigue el 70% de sus puntos en tiros de media distancia desde el lateral derecho.

En general, disponer de un modelo que permita simular el comportamiento y/o predecir la evolución de un sistema, un proceso, las ventas de un producto, etc. de forma suficientemente precisa supone una clara ventaja competitiva, permitiendo adelantarse y aprovechar oportunidades, así como prevenir problemas. Algunas de las aplicaciones más comunes son:

3.2.1 Marketing

Este es uno de los campos donde los éxitos de la minería de datos son más conocidos. Cuanto más precisa sea la información que tengamos sobre los clientes, mayores posibilidades tendremos de aumentar nuestros ingresos y rentabilizar al máximo nuestras acciones. El objetivo fundamental puede resumirse en determinar *quién comprará qué, cuándo y dónde*.

- ✓ Targeting: Podemos aumentar espectacularmente el porcentaje de respuesta a una campaña de marketing si se dirige a los objetivos adecuados. La minería de datos permite detectar entre los potenciales clientes los que presentan una mayor probabilidad de responder a la campaña y dirigirla a ellos específicamente, con lo cual se consigue reducir drásticamente los costes.

- ✓ Fidelización de clientes: Conseguir un nuevo cliente o recuperar uno perdido resulta mucho más costoso que mantener uno que ya lo es. De ahí la rentabilidad de las campañas de fidelización de clientes, que detectan aquéllos que parece más probable que se vayan a perder, permitiendo llevar a cabo iniciativas que eviten dicha pérdida.

- ✓ La minería de datos también permite detectar nuevas oportunidades de mercado, comparando hábitos de consumo de diferentes clientes, por ejemplo, o determinando la ubicación más conveniente para un determinado negocio.

3.2.2 Predicción

Conocer a priori cómo evolucionará una variable en el futuro constituye una información muy valiosa y supone una indudable ventaja competitiva. Se trata de una herramienta de evidente interés tanto desde el punto de vista comercial, como en gestión o control de procesos.

A partir de los datos históricos almacenados y utilizando técnicas de minería de datos pueden elaborarse modelos que permitan estimar con precisión la evolución de una variable en el futuro. Disponer de esta información con tiempo suficiente permite adecuar la respuesta de forma óptima. Esto puede resultar útil en los campos más diversos:

- ✓ Detección de oportunidades.
- ✓ Prevención de problemas.
- ✓ Gestión óptima del personal.
- ✓ Optimización de stocks.

3.2.3 Reducción de riesgos

La minería de datos permite construir sistemas de evaluación automática de riesgos, basados en la experiencia previa. Estos sistemas resultan de gran utilidad cuando la cantidad de casos a evaluar es excesiva para su procesamiento manual. El empleo de técnicas de minería de datos ha aumentado la eficacia y fiabilidad de dichos sistemas, logrando un comportamiento más similar al de los expertos humanos.

3.2.4 Detección de fraudes

Aplicando técnicas de minería de datos, pueden obtenerse modelos que permitan descubrir posibles fraudes, basándose en la detección de comportamientos anómalos, en comparación con los datos registrados anteriormente.

Podemos encontrar aplicaciones concretas en operadores de telefonía o empresas de gestión de tarjetas de crédito. Estas compañías analizan el uso que los clientes hacen de sus servicios y pueden localizar, de manera muy rápida, un uso fraudulento de los mismos.

3.2.5 Control de calidad

Existen numerosos ejemplos en los que se han aplicado técnicas de minería de datos para desarrollar sistemas automáticos de control de calidad. Estos sistemas suponen un considerable ahorro en el proceso productivo, puesto que facilitan:

- ✓ ***Detección más precisa de productos defectuosos***
  A menudo el control de calidad se realiza de forma manual y, por tanto, depende de una evaluación subjetiva por parte del personal encargado del mismo. El principal problema de este método es que el criterio de calidad no es estable sino que depende de la persona que realiza el análisis. La minería de datos permite desarrollar sistemas automáticos de control de calidad que discriminan los productos defectuosos con un alto grado de precisión y fiabilidad, según un criterio objetivo.
  Esto no sólo evita el problema mencionado anteriormente. Además, al aumentar la exactitud de la evaluación se ahorran los costes derivados de las



clasificaciones erróneas: productos defectuosos que se consideraron correctos por error y productos correctos, desechados por un exceso de precaución.

- ***Localización precoz de defectos***

  El control de calidad no sólo debe realizarse al final del proceso. Cuanto antes se detecte un fallo, menor será su impacto. Además de las ventajas de los sistemas automáticos ya comentadas, en este caso existe un problema añadido. A menudo no resulta fácil medir la variable que determina la calidad del producto en tiempo real o en la cadena de producción. En estos casos, es imprescindible utilizar técnicas de minería de datos para descubrir posibles relaciones que permitan detectar los fallos utilizando las variables disponibles durante el proceso.

- ***Identificación de causas de fallos***

  La minería de datos no sólo resulta útil para discriminar los productos defectuosos. También ayuda a determinar los fallos más frecuentes así como identificar las causas de los mismos. Esto permite adoptar medidas para evitarlos en el futuro.

- ***Análisis no destructivo***

  A menudo, para obtener la información que se necesita, hay que realizar un análisis destructivo. Un ejemplo típico es la evaluación de la resistencia de un material, medida que se establece forzándolo hasta que se rompe. Utilizando minería de datos es posible estimar con bastante exactitud el valor de este tipo de parámetros en función de otras características que sí pueden medirse sin destruir el producto. Esto permite controlar la calidad de todos los productos fabricados y no sólo de una pequeña muestra, ya que no se destruyen con el examen.

3.2.6 Procesos industriales

Otra aplicación básica de la minería de datos en el entorno industrial, además del control de calidad, es el control de procesos. Estas técnicas permiten explotar la información disponible sobre un sistema o proceso y utilizar los modelos desarrollados (bien de un sistema o proceso global, o bien de una parte concreta del mismo) para:

- ***Automatizar y optimizar el control del proceso***

  En muchos sistemas se conoce el proceso suficientemente como para diseñar e implantar controladores a partir de análisis matemático del proceso. En otras ocasiones, esto no es posible, bien por que el proceso es enormemente complejo, bien porque no disponemos de todas las variables. En estas circunstancias, técnicas de minería de datos pueden ayudarnos a establecer relaciones entre las variables, y así diseñar los controladores adecuados.

- ***Optimizar su rendimiento***

  Los propios sistemas de aprendizaje pueden ser utilizados para adaptar los mecanismos de control de forma permanente, en función de los datos del proceso que vayamos recibiendo. De esta forma es posible optimizar el rendimiento del proceso, adaptando los controladores, en cada momento, a la situación de la planta.

- ***Implementar programas de mantenimiento predictivo***

  Uno de los problemas de todo equipo de mantenimiento de un proceso es establecer el calendario de reparaciones. Las reparaciones, limpiezas y ajustes programados suponen en muchos casos parar el proceso productivo, con las consiguientes pérdidas, no sólo de lo que se deja de producir sino de los costes de parada y arranque de la cadena. Un análisis profundo de los datos de que se disponga puede permitir hacer una planificación óptima de estas paradas, de manera que se minimice su impacto.

4 CONCLUSIONES

La minería de datos es una herramienta que permite convertir los datos recogidos durante el funcionamiento normal de nuestro negocio en información valiosa. No es una tecnología que suplante a otras, sino que es complementaria y, en muchos casos, se aprovecha de lo que otros mecanismos, como la estadística, puedan aportarle.

Técnicas como el agrupamiento y la clasificación automática de clientes facilitan el diseño y puesta en marcha de planes de marketing mucho más eficaces. Si nuestro trabajo se centra en el entorno industrial, la minería de datos puede aportar información valiosa sobre la calidad de nuestros productos, el mantenimiento preventivo o la propia optimización de nuestros procesos. Si nos movemos en las nuevas tecnologías, el análisis del acceso a nuestros servidores de internet, puesto en relación con las ventas realizadas o los servicios ofrecidos, será más potente utilizando *web mining* que haciendo un simple análisis de tráfico. En resumen, la minería de datos nos permite tomar una posición en nuestro mercado que nos diferencie de nuestros competidores.

DAEDALUS-Data, Decisions and Language, S.A. pone al servicio de sus clientes la experiencia de sus profesionales en el aprendizaje automático, la minería de datos y los servicios telemáticos durante más de 10 años. Una experiencia que nos permite afrontar los nuevos retos tecnológicos con la mayor seguridad.

